# Anonymizing Radiographs Using an Object Detection Deep Learning Algorithm

*Bardia Khosravi, MD, MPH, MHPE** • *John P. Mickley, MD** • *Pouria Rouzrokh, MD, MPH, MHPE* • *Michael J. Taunton, MD* • *A. Noelle Larson, MD* • *Bradley J. Erickson, MD, PhD* • *Cody C. Wyles, MD*

From the Orthopedic Surgery Artificial Intelligence Laboratory, Department of Orthopedic Surgery (B.K., J.P.M., P.R., M.J.T., A.N.L., C.C.W.), Radiology Informatics Laboratory, Department of Radiology (B.K., P.R., B.J.E.), Department of Orthopedic Surgery (M.J.T., A.N.L., C.C.W.), and Department of Clinical Anatomy (C.C.W.), Mayo Clinic, 200 1st St SW, Rochester, MN 55905. Received March 19, 2023; revision requested April 13; revision received August 11; accepted August 25. **Address correspondence to** C.C.W. (email: *Wyles.Cody@mayo.edu*).

Supported by the Mayo Foundation Presidential Fund.

* B.K. and J.P.M. contributed equally to this work.

Conflicts of interest are listed at the end of this article.

See also the commentary by Chang and Li in this issue.

*Radiology: Artificial Intelligence* 2023; 5(6):e230085 • https://doi.org/10.1148/ryai.230085 • Content code: AI

Radiographic markers contain protected health information that must be removed before public release. This work presents a deep learning algorithm that localizes radiographic markers and selectively removes them to enable de-identified data sharing. The authors annotated 2000 hip and pelvic radiographs to train an object detection computer vision model. Data were split into training, validation, and test sets at the patient level. Extracted markers were then characterized using an image processing algorithm, and potentially useful markers (eg, "L" and "R") without identifying information were retained. The model achieved an area under the precision-recall curve of 0.96 on the internal test set. The de-identification accuracy was 100% (400 of 400), with a de-identification false-positive rate of 1% (eight of 632) and a retention accuracy of 93% (359 of 386) for laterality markers. The algorithm was further validated on an external dataset of chest radiographs, achieving a de-identification accuracy of 96% (221 of 231). After fine-tuning the model on 20 images from the external dataset to investigate the potential for improvement, a 99.6% (230 of 231, $P = .04$) de-identification accuracy and decreased false-positive rate of 5% (26 of 512) were achieved. These results demonstrate the effectiveness of a two-pass approach in image de-identification.

*Supplemental material is available for this article.*



Radiographs are the most frequently acquired radiologic studies, given their low cost and low radiation dose (1). Most radiographs have radiopaque markers that display the side, positioning, radiographer initials, date, and patient identifier. Marker placement is typically done by technicians, resulting in substantial variation between studies (2).

Deep learning (DL) algorithms require high volumes of diverse data to achieve optimal performance. One approach is to pool data from multiple institutions to achieve a sufficiently diverse population to train a DL model. However, such studies require a Health Insurance Portability and Accountability Act waiver, which often requires that all research data be de-identified (3,4). Radiographic markers present a major challenge to the de-identification process as they are a part of the image and are not accessible by simply filtering image metadata (5). Also, these markers potentially provide models with a false path to cut corners when reaching an algorithmic conclusion (6). Therefore, removing these markers is a crucial preprocessing step before data sharing, which is labor intensive.

In this work, we developed a DL algorithm to localize and remove radiologic markers, enabling de-identified data sharing and also making the images more suitable for robust DL training. Additionally, we implemented safeguards to retain laterality markers as an example to demonstrate the model's adaptability to retain desired information. We further assessed its generalizability on an external, out-of-domain dataset.

## Materials and Methods

### Image Dataset and Model Development

This study was approved by the institutional review board with a waiver of informed consent. We annotated 2000 native anteroposterior pelvic radiographs, randomly selected from an institutional hip arthroplasty registry, which were obtained between January 2000 and January 2021 (7). For this study, we were interested in all markers that were related to the patient, type of examination, institution, and staff. One annotator (J.P.M.) with 4 years of experience annotated these markers by drawing a bounding box around them (4). Then, the data were split at the patient level into 60%/20%/20% for training, validation, and testing, respectively. Images were resampled to 512 × 512 pixels, and a YOLOv5-x (Ultralytics) model was trained to detect the marker location (8,9). More details can be found in Appendix S1 and at *https://github.com/OrthopedicSurgeryAILab/RadiographMarkerRemover*.

Although removing identification markers is crucial, other markers may be desirable for retention for preprocessing steps. The prevalence of laterality markers in our data made them an excellent choice to demonstrate how desired markers can be retained (10). However, it has been shown that even these markers can lead to shortcut learning and should therefore be redacted in the training phase (11). After a review of internal and external datasets, such as CheXpert, we developed a postprocessing algorithm to

### Abbreviations

AUPRC = area under the precision-recall curve, DL = deep learning, mAP-50 = mean average precision at a 50% threshold, OCR = optical character recognition


### Summary

The proposed deep learning model was able to robustly detect and remove radiographic markers from radiographs of different anatomic regions, yielding anonymized images suitable for the development of artificial intelligence algorithms.


### Key Points

- A two-pass algorithm approach was used to localize and characterize radiologic markers for image anonymization.
- Fine-tuning the localizer network increased de-identification accuracy to 99.6% ($P$ = .04).
- Selective retention of markers by the deep learning model enables granular control over image de-identification.


### Keywords

Conventional Radiography, Skeletal-Axial, Thorax, Experimental Investigations, Supervised Learning, Transfer Learning, Convolutional Neural Network (CNN)


retain these markers (12). After localizing all markers, the algorithm analyzes each detected marker through a series of steps. First, the Otsu method is applied to find the threshold, converting the marker area into a binary image. The largest connected area within this binary image is then identified. This component undergoes an automatic correction to make the marker's orientation upright. Finally, an optical character recognition (OCR) algorithm equipped with long short-term memory examines the component to detect isolated R and L characters. If identified as either of these characters, the component is kept in the original image; if not, it is replaced by black pixels (Fig 1). It should be mentioned that these operations were carried out on original-resolution images.

Annotations were created using LabelMe software version 5.0.1 (13), and models were trained using PyTorch (version 1.11.0). Character recognition algorithms from Tesseract (version 5.0.0) were used for marker retention (14).

### Model Evaluation and Statistical Analysis

The mean average precision at a 50% threshold (mAP-50) and area under the precision-recall curve (AUPRC) were reported as measures of localization model performance. As an image-level metric, de-identification accuracy was defined as the ability to remove *all* markers from a radiograph other than the L and R markers. De-identification false-positive rate was used to show the rate of incorrectly detected bounding boxes. Retention accuracy reflects the number of laterality markers that were successfully detected and retained on the image.

For external testing, we used the validation set of the CheXpert dataset ($n$ = 234) to evaluate the model's performance on data from different organizations and anatomic regions (12). Three images from this set were excluded from the analysis, as they did not have any markers. Through backbone freezing, the model was further fine-tuned on 20 randomly selected images from the CheXpert training set to investigate the potential for improvement in target task performance (14).

Overall pipeline speed is reported for both the CPU (i7 2.5 GHz; Intel) and GPU (V100; NVIDIA), averaged over 100 images. Fisher exact test from the SciPy package (version 1.10) was used to measure performance improvement, with a significance level of .05.

## Results

The model achieved an AUPRC of 0.98 (mAP-50: 0.99) on the validation set and 0.96 (mAP-50: 0.98) on the internal test set of pelvic radiographs, with a de-identification accuracy of 400 of 400 (100%; Fig 2) on the test set. The de-identification false-positive rate was 1% (eight of 632). Retention accuracy was 93% (359 of 386). The whole pipeline took 1.43 seconds per image on the CPU and 0.51 second per image on the GPU because of the faster execution of the localization model.

On the external test set, the model achieved a de-identification accuracy of 96% (221 of 231). After fine-tuning, the model's de-identification accuracy improved to 99.6% (230 of 231) ($P$ = .04, Fig 2). The de-identification false-positive rate of the fine-tuned model was 5% (26 of 512), compared with 11% (54 of 508) before fine-tuning. Retention accuracy was 91% (211 of 231) after fine-tuning. The Table provides a summary of the results. Most of the false-positive boxes were electrocardiography leads that were not seen during training (Fig 3).

## Discussion

Radiographs are a routine part of many diagnostic workflows and are very popular for developing DL algorithms. However, the presence of burnt-in identifiers on these images causes several obstacles in model training. First, most research data sharing agreements require a Health Insurance Portability and Accountability Act waiver, meaning all protected health information must be removed. Radiographic markers limit data sharing and the creation of public, multicentric datasets. Second, due to their high intensity, radiographic markers introduce spectrum bias (ie, widely discrepant normalization results when comparing images with and without markers), which can distract the algorithm (15). Finally, it has been shown that DL models can detect differences in hospital systems, which may reflect inpatient or outpatient settings, using them as a proxy for patient acuity (11). For instance, radiographs taken with emergency department devices are more likely to show a fracture than those from outpatient clinics, impacting the pretest probability (6). Hence, being able to strip this information from radiographs can help with more robust model training. Our proposed model demonstrated excellent de-identification accuracies of 100% on the internal test set and 99.6% on the external test set after model fine-tuning, while retaining laterality markers.

There have been several efforts to tackle pixel-level image de-identification (16). Newhauser et al (17) showed that thresholding worked better than OCR to remove text pixels from radiation therapy planning images. As mentioned, the problem with OCR algorithms is that they need clean, preprocessed images

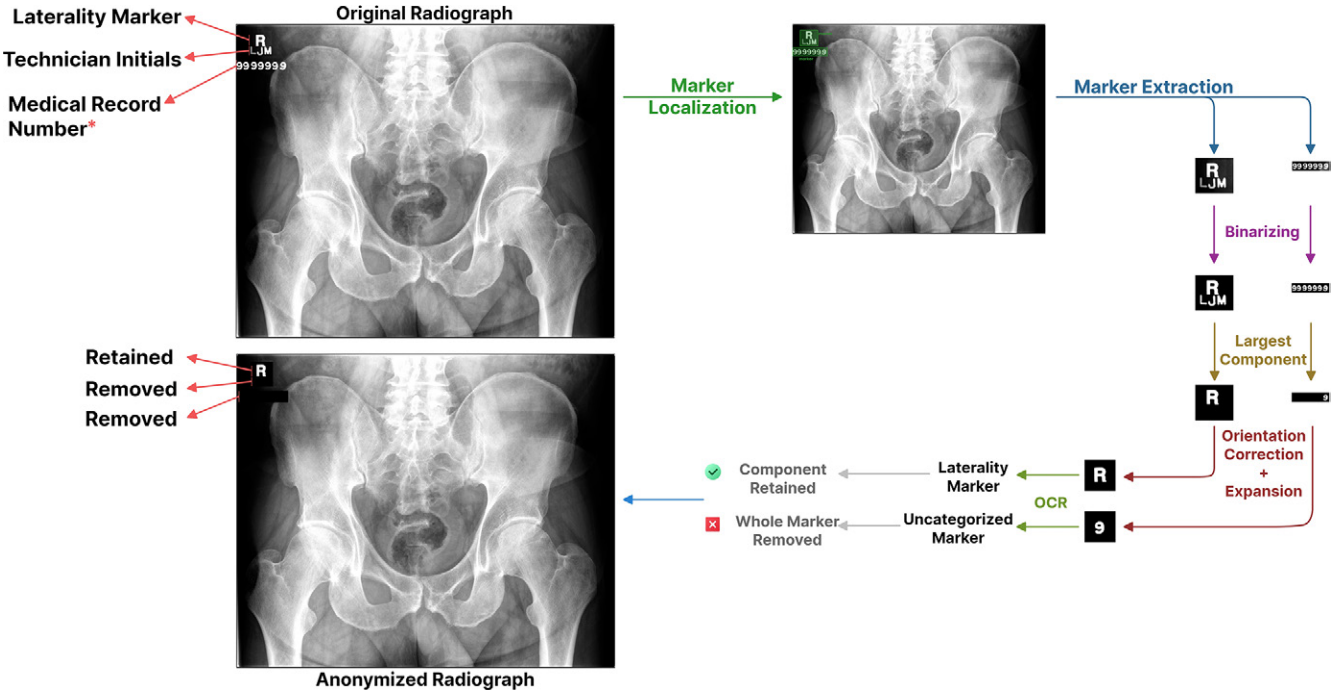


**Figure 1:** Overview of the proposed pipeline. * The patient identifiers were in the same place, but they were taken out and replaced with fake numbers, using the same font settings.

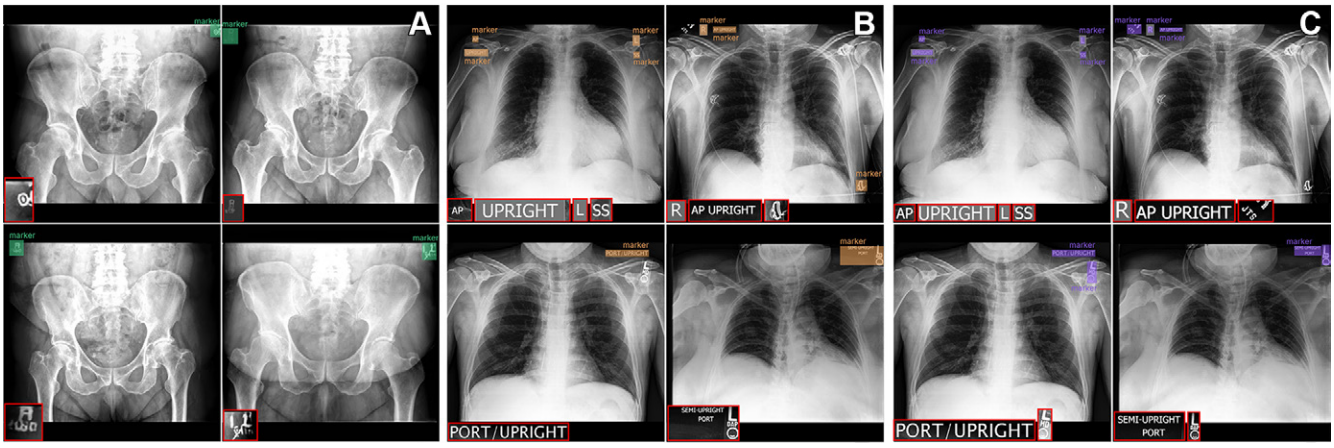


**Figure 2:** **(A)** Marker detection on pelvic radiographs. Note that the marker areas were obfuscated to ensure patient privacy. **(B, C)** Results of the localizer model's performance on out-of-domain chest radiographs before **(B)** and after **(C)** fine-tuning. The red boxes are zoomed-in versions of the marker areas that were detected. As shown, the fine-tuned model draws tighter bounding boxes with fewer nonmarker areas and also has more specificity, not mistaking nonmarker densities, like electrocardiography leads, compared with the initial model.

and can have varying performances, ranging from 89% to 98% recall in terms of annotation detection (18). Moreover, versatility of the markers poses a problem for conventional OCR algorithms. However, a one-pass DL algorithm missed 10% of medical images with identifiable patient information (19). Our method could be complemented with obfuscation of biometric information in the radiographs to further ensure patient privacy and prevent reidentification (20).

Our approach focused on a two-pass marker removal, as follows: In the first pass, a highly sensitive algorithm is run to detect all marker areas to prevent protected health information leak; in the second pass, an OCR algorithm is used to allow informative tags, like laterality markers, to bypass redaction. This modular approach also facilitates adding filters to find and stop the removal of certain markers, such as those that are specific to an institution. Moreover, due to the large variability between different institutions in the types and use cases of imaging markers, having an easily adaptable approach that can be fine-tuned with minimal effort is desirable. As we showed, adding only 20 images from the target domain greatly improved generalizability. Finally, the second pass can be disabled while training a DL model to prevent shortcut learning based on burnt-in markers.

Our results should be interpreted with some limitations. First, while we had perfect results on our internal test sets, all the external public images had their protected health information removed, which forced us to evaluate our results on

**Summary of Model Accuracy Results**

| Metric | Internal Test Set | External Test Set (CheXpert) | External Test Set (CheXpert) Fine-tuned |
|---|---|---|---|
| De-identification accuracy (%) | 100.0 (400/400) | 95.8 (221/231) | 99.6 (230/231) |
| De-identification false-positive rate (%) | 1.2 (8/632) | 10.6 (54/508) | 5.0 (26/512) |
| Retention accuracy (%) | 93.0 (359/386) | 87.9 (203/231) | 91.3 (211/231) |

Note.—De-identification accuracy is calculated as the number of radiographs with all markers removed divided by the total number of radiographs. De-identification false-positive rate is calculated as the number of false-positive bounding boxes divided by the total number of bounding boxes. Retention accuracy is calculated as the number of retained laterality markers divided by the total number of laterality markers.

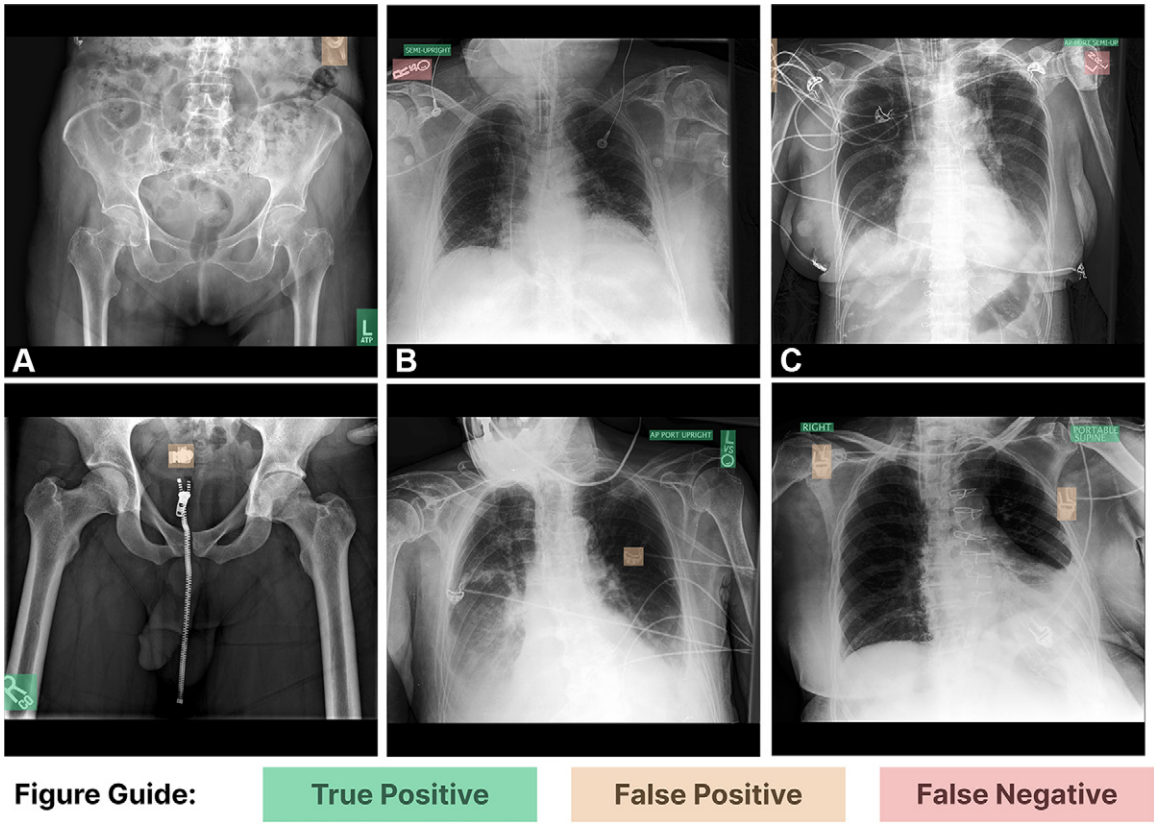


**Figure 3:** Cases of model failure on the **(A)** internal and external datasets **(B)** before and **(C)** after fine-tuning.

the study- and staff-related markers. Second, we developed the marker retention algorithm on the basis of the most frequent type of markers that we found in our survey, but the algorithm might need fine-tuning to cover other types of markers. Finally, prior to broader deployment, there is still a need for more studies on other body parts and modalities to ensure the maintenance of robust algorithm performance.

In summary, we developed a model that can accurately and selectively remove radiographic markers from radiographs and has built-in flexibility for iterative improvement. We anticipate that this model can be easily refined to effectively detect radiographic markers on a wide variety of radiographs, highlighting its generalizability.

**Author contributions:** Guarantors of integrity of entire study, **B.K., A.N.L., C.C.W.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **B.K., J.P.M., A.N.L., C.C.W.**; clinical studies, **M.J.T., A.N.L., C.C.W.**; experimental studies, **B.K., J.P.M., B.J.E., C.C.W.**; statistical analysis, **B.K., J.P.M., P.R., B.J.E.**; and manuscript editing, all authors

**Disclosures of conflicts of interest:** **B.K.** Member of the *Radiology: Artificial Intelligence* trainee editorial board. **J.P.M.** No relevant relationships. **P.R.** Member of RadioGraphics TEAM. **M.J.T.** No relevant relationships. **A.N.L.** Royalties from Globus to Mayo Orthopedics and to author; consulting fees from Globus, Orthopediatrics, Stryker, Depuy Synthes, Alexion, ZimVie, and Medtronic (all funds directed to Mayo orthopedic research); patent to Mayo on vertebral body tethering; participation on Medtronic Advisory Board and Steering Committee for Braive study; board member at POSNA and the Setting Scoliosis Straight Foundation; committee member at the Scoliosis Research Society; leadership or fiduciary role on the research council at the Pediatric Spine Study Group. **B.J.E.** Research chair for SIIM; stock/stock options in FlowSIGMA, VoiceIT, and Yunu; consultant to the editor for *Radiology: Artificial Intelligence*. **C.C.W.** No relevant relationships.

## References


1. National Health Services Government Statistical Service (NHS-GSS). Diagnostic Imaging Dataset 2021-22 Data. https://www.england.nhs.uk/statistics/statistical-work-areas/diagnostic-imaging-dataset/diagnostic-imaging-dataset-2021-22-data/. Published 2023. Accessed January 4, 2023.
2. Barry K, Kumar S, Linke R, Dawes E. A clinical audit of anatomical side marker use in a paediatric medical imaging department. J Med Radiat Sci 2016;63(3):148–154.
3. Edwards A, Hollin I, Barry J, Kachnowski S. Barriers to cross--institutional health information exchange: a literature review. J Healthc Inf Manag 2010;24(3):22–34.
4. Institute of Medicine · Board on Health Care Services · Board on Health Sciences Policy · Committee on Health Research and the Privacy of Health Information: The HIPAA Privacy Rule. Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research. National Academies Press; 2009. https://play.google.com/store/books/details?id=jUGcAgAAQBAJ. Accessed February 1, 2023.
5. Rutherford M, Mun SK, Levine B, et al. A DICOM dataset for evaluation of medical image de-identification. Sci Data 2021;8(1):183.
6. Badgeley MA, Zech JR, Oakden-Rayner L, et al. Deep learning predicts hip fracture using confounding patient and healthcare variables. NPJ Digit Med 2019;2(1):31.
7. Rouzrokh P, Khosravi B, Johnson QJ, et al. Applying deep learning to establish a total hip arthroplasty radiography registry: a stepwise approach. J Bone Joint Surg Am 2022;104(18):1649–1658.
8. Jocher G, Chaurasia A, Stoken A, et al. ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference. Published February 22, 2022. Accessed February 1, 2023.
9. Touvron H, Vedaldi A, Douze M, Jégou H. Fixing the train-test resolution discrepancy. arXiv1906.06423 [preprint] https://arxiv.org/abs/1906.06423. Posted June 14, 2019. Accessed February 1, 2023.
10. DeGrave AJ, Janizek JD, Lee SI. AI for radiographic COVID-19 detection selects shortcuts over signal. medRxiv 2020.09.13.20193565 [preprint] Posted October 8, 2020. Accessed February 1, 2023.
11. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med 2018;15(11):e1002683.
12. Irvin J, Rajpurkar P, Ko M, et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. arXiv 1901.07031 [preprint] https://arxiv.org/abs/1901.07031. Posted January 21, 2019. Accessed February 1, 2023.
13. LabelMe software. wkentaro/labelme: v4.6.0. Zenodo. 2021. https://doi.org/10.5281/zenodo.5711226. Updated November 18, 2021. Accessed February 1, 2023.
14. Zhuang F, Qi Z, Duan K, et al. A comprehensive survey on transfer learning. Proc IEEE 2021;109(1):43–76.
15. Kim KD, Cho K, Kim M, et al. Enhancing deep learning based classifiers with inpainting anatomical side markers (L/R markers) for multi-center trials. Comput Methods Programs Biomed 2022;220:106705.
16. Aryanto KYE, Oudkerk M, van Ooijen PMA. Free DICOM de-identification tools in clinical research: functioning and safety of patient privacy. Eur Radiol 2015;25(12):3685–3695.
17. Newhauser W, Jones T, Swerdloff S, et al. Anonymization of DICOM electronic medical records for radiation therapy. Comput Biol Med 2014;53:134–140.
18. Vcelak P, Kryl M, Kratochvil M, Kleckova J. Identification and classification of DICOM files with burned-in text content. Int J Med Inform 2019;126:128–137.
19. Monteiro E, Costa C, Oliveira JLA. A de-identification pipeline for ultrasound medical images in DICOM format. J Med Syst 2017;41(5):89.
20. Packhäuser K, Gündel S, Thamm F, Denzinger F, Maier A. Deep learning-based anonymization of chest radiographs: a utility-preserving measure for patient privacy. arXiv 2209.11531 [preprint] https://arxiv.org/abs/2209.11531. Posted September 23, 2022. Accessed February 1, 2023.